Genotyping Structural Variations using Long Read Data

Lolita Lecompte[1] Pierre Peterlongo[1] Dominique Lavenier[1] Claire Lemaitre[1]

[1]Univ Rennes, CNRS, Inria, IRISA - UMR 6074, F-35000 Rennes, France

Background

Structural variations (SV) are characterized as genomic segments of a least 50 base pairs (bp) long, that are rearranged in the genome. There are several types of SV such as deletions, insertions, duplications, inversions, translocations. This kind of polymorphism have been shown involved in many biological processes, particularly diseases or evolution [1]. Databases referencing such variants grow as new variants are discovered, at this time dbVar, the reference database of human genomic SVs [2], contains 35,428,724 variant calls, illustrating that many SVs have already been discovered and characterized in the human population. In this context, it becomes very interesting and informative to evaluate for a given newly sequenced individual if its genome holds already known SVs. This is commonly known as the SV genotyping problem.

Such genotyping methods already exist for short reads data: for instance, SVtyper [3], SV^2 [4]. Though short reads are often used to discover and genotype SVs, this is well known that their short size make them ill-adapted for predicting large SVs or SVs located in repeated regions. Third generation sequencing technology, such as Pacific Biosciences (PB) and Oxford Nanopore Technologies (ONT), can produce long reads data compared to Next Generation Sequencing technologies. Despite their higher error rate, long reads are crucial in the study of SVs. Indeed, the size range of this data can reach a few kilobases to megabases, thus long reads can extend over rearranged SV sequences as well as over the repeated sequences often present at SV's breakpoint regions.

Following long reads technology's development, many SV discovery tools have emerged, such as Sniffles [5]. To our knowledge there is currently no tool that can perform genotyping from a set of known SVs with long reads data. Thus, there is a need to develop accurate and efficient methods to genotype SVs with long reads data, especially in the context of clinical diagnoses.

Results

Method We propose a novel method that aims at assigning a genotype for a set of already known SVs in a given individual sample sequenced with long reads data. In other words, the method assesses if each SV is present in the given individual, and if so, how many variant alleles it holds, ie. whether the individual is heterozygous or homozygous for the particular variant. The method is described and implemented here for only one type of SV, the deletions, but the principle can be easily generalized to other types of SVs. We also provide an implementation of this method in the tool named SVJedi.

The principle of the method is based on: **1)** Generating reference sequences that represent the two possible alleles of each SV. The reference allele (allele 0) is therefore the sequence of the deletion with adjacent sequences at each side, and the alternative allele (allele 1) consists in the joining of the two previous adjacent sequences. **2)** Then, sequenced long reads are aligned on all previously generated references, using Minimap2 [6], specifically designed for long erroneous reads. **3)** An important step of our method consists in selecting informative alignments, in order to remove i) uninformative alignments, that is those not discriminating between the two possible alleles, and ii) spurious false positive alignments, that are mainly due to repeated sequences. **4)** Finally, for each SV, the allele frequency is measured based on the number of supporting alignments, in order to estimate genotype.

Evaluation on simulated data SVJedi was assessed on PB simulated long reads for the human chromosome 1, with 1,000 real characterized deletions found in dbVar [2], ranging from 50 to 10,000 bp, equally distributed among the three different genotypes (0/0, 0/1, 1/1).

SVJedi achieved 95.8 % precision, it correctly assigned genotypes to 942 over 987 predicted deletions. Most erroneous genotypes concern deletions of small size (less than 100 bp), as expected these are harder to genotype than longest deletions. As a matter of fact, the precision is of 85.4 % for deletions smaller than 100 bp versus 97.9 % for deletions greater than 500 bp. The remaining false positive deletions of size $\geq$ 100 bp, were manually investigated, and most of them occur in regions with a high density of mobile elements.

Comparison with SV discovery approaches Then we assessed if these simulated deletions could be easily detected and genotyped by a long read SV discovery tool. We applied here the best to date such tool, Sniffles [5] to the chromosome 1 simulated read dataset. As expected, none of the 333 simulated deletions with 0/0 genotypes were assigned a genotype in the Sniffles output call set, since a discovery tool naturally only reports present variants. Surprisingly, among the 667 deletions simulated with either a 0/1 or 1/1 genotype, only 406 were discovered by Sniffles, which gives a recall of only 60.9 %. Interestingly, Sniffles also mis-predicts the genotype of the discovered deletions, assigning most of the 1/1 discovered deletions (n = 254, 81 %) as heterozygous. This highlights the fact that Sniffles, a SV discovery tool, is much less precise for the genotyping task than a dedicated genotyping tool.

Application to real human data SVJedi was also applied on real ONT data [7] for the whole human genome of NA12878 individual. As the set of deletions to genotype, we used the merged SV call set provided by the Genome in a Bottle (GiAB) consortium for NA12878 (Mt. Sinai School of Medicine dataset), where only SVs predicted by all methods were kept. This set of known variants contains 1,685 deletions.

SVJedi assigned a genotype to 1,684 deletions, of which 1,514 (90 %) were genotyped exactly as in GiAB. SVJedi took 1h46m on this dataset, including 1h42 for the alignment with Minimap2 parallelized on 40 cpu and with a maximum RAM memory of 6.5 Gbytes.

Conclusions

In this work, we provide a novel SV genotyping approach for long reads data, that is fast and accurate on both simulated and real datasets. This work also demonstrated that this is crucial to develop dedicated SV genotyping methods, since SV discovery methods are not as efficient and precise to genotype variants once SVs have been discovered. The approach is implemented for the moment only for deletion variants in the SVJedi software. However, this proof of principle on deletion variants is a first step before generalizing the approach for all types of SVs. Insertion variants are simply the counterpart of deletions, and inversions and translocations are SVs even more balanced than insertions/deletions regarding the number of breakpoints (with exactly two breakpoints per allele). Therefore, for all these types of SVs, the method will be easily generalized. Our method fills a gap and now enables SV genotyping using long reads for clinical diagnosis or population genotyping. SVJedi is available at `https://data-access.cesgo.org/index.php/s/6EhOdOBsVNRr72n`, under GNU Affero GPL licence.

References

[1] James R Lupski. Structural variation mutagenesis of the human genome: Impact on disease and evolution. *Environmental and molecular mutagenesis*, 56(5):419–436, 2015.

[2] Lon Phan, Jeffrey Hsu, Le Quang Minh Tri, Michaela Willi, Tamer Mansour, Yan Kai, John Garner, John Lopez, and Ben Busby. dbvar structural variant cluster set for data analysis and variant comparison. *F1000Research*, 5, 2016.

[3] Colby Chiang, Ryan M Layer, Gregory G Faust, Michael R Lindberg, David B Rose, Erik P Garrison, Gabor T Marth, Aaron R Quinlan, and Ira M Hall. Speedseq: ultra-fast personal genome analysis and interpretation. *Nature methods*, 12(10):966, 2015.

[4] Danny Antaki, William M Brandler, and Jonathan Sebat. Sv2: accurate structural variation genotyping and de novo mutation detection from whole genomes. *Bioinformatics*, 34(10):1774–1777, 2017.

[5] Fritz J Sedlazeck, Philipp Rescheneder, Moritz Smolka, Han Fang, Maria Nattestad, Arndt von Haeseler, and Michael C Schatz. Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods*, 15(6):461–468, 2018.

[6] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 2018.

[7] Miten Jain, Sergey Koren, Karen H Miga, Josh Quick, Arthur C Rand, Thomas A Sasani, John R Tyson, Andrew D Beggs, Alexander T Dilthey, Ian T Fiddes, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature biotechnology*, 36(4):338, 2018.